THE HYPERBOLIC LATTICE POINT COUNT IN INFINITE VOLUME WITH APPLICATIONS TO SIEVES

ALEX V. KONTOROVICH

ABSTRACT. We develop novel techniques using abstract operator theory to obtain asymptotic formulae for lattice counting problems on infinite-volume hyperbolic manifolds, with error terms which are uniform as the lattice moves through "congruence" subgroups. We give the following application to the theory of affine linear sieves. In the spirit of Fermat, consider the problem of primes in the sum of two squares, $f(c, d) = c^2 + d^2$, but restrict (c, d) to the orbit $\mathcal{O} = (0, 1)\Gamma$, where Γ is an infinite-index non-elementary finitely-generated subgroup of $\mathrm{SL}(2, \mathbb{Z})$. Assume that the Reimann surface $\Gamma \backslash \mathbb{H}$ has a cusp at infinity. We show that the set of values $f(\mathcal{O})$ contains infinitely many integers having at most R prime factors for any $R > 4/(\delta - \theta)$, where $\theta > 1/2$ is the spectral gap and $\delta < 1$ is the Hausdorff dimension of the limit set of Γ. If $\delta > 149/150$, then we can take $\theta = 5/6$, giving $R = 25$. The limit of this method is $R = 9$ for $\delta - \theta > 4/9$. This is the same number of prime factors as attained in Brun's original attack on the twin prime conjecture.

CONTENTS

1. Introduction

Many of the most enticing problems in number theory amount to finding primes or almost-primes (numbers having few prime factors) in "thin" subsets of the integers. By thin, we mean that the number of elements in the given set of size not exceeding a height T is bounded by T^δ, where δ is some constant less than 1. See the papers [Pu53, Che73, Iwa78, FI98, HB01] which are landmarks in sieve theory, producing primes or products of at most two primes in thin sets.

It is our main goal to exhibit such a set arising from the orbit of an affine linear group action and sift for elements having few prime factors. Our starting point is the fundamental work of Bourgain, Gamburd, and Sarnak [BGS06, BGS08, Sar07] on the so-called affine linear sieve:

Theorem 1.1 ([BGS06]). *Let $\Gamma \subset SL_2(\mathbb{Z})$ be any non-elementary[1] subgroup, let $q \in \mathbb{Z}^2$ be nonzero, and let $\mathcal{O} = q \cdot \Gamma$ be a Γ-orbit. Let $f : \mathbb{Z}^2 \to \mathbb{Z}$ be any polynomial. Then there exists an $R < \infty$, depending on all of the above data, such that there are infinitely many points in the set $f(\mathcal{O})$ having at most R prime factors.*

Notice that there are no congruence conditions, nor conditions on f (such as irreducibility) – these are all factored into R, which is left completely unspecified. There are three main ingredients:

(1) Number Theory – the combinatorial sieve,
(2) Algebra – Strong Approximation and Goursat's Lemma, and
(3) Combinatorics – counting by wordlength in Γ and extending the recently established expander property of Bourgain-Gamburd [BG07] (with an unspecified spectral gap) to square-free moduli.

The lack of specificity of the spectral gap (and arbitrariness of the choice of orbit $\mathcal{O}$ and function f) induces a lack of specificity of the number R of prime factors.

In this paper, we select a particular orbit $\mathcal{O}$ and function f, and give a precise bound for the number R of prime factors by replacing input (3) above with an archimedian count and using Gamburd's explicit 5/6-th gap [Gam02] in place of the unspecified spectral gap.

This amounts to a hyperbolic lattice point counting problem, requiring uniform error estimates as the lattice moves through "congruence" subgroups. In the interest of having a thin set, we are compelled to

---

[1] Recall that an elementary group $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ has the property that any two elements $g, h \in \Gamma$ of infinite order have $|\mathrm{tr}(ghg^{-1}h^{-1})| = 2$, see [Bea83]. These are the "abelian" groups, such as purely hyperbolic groups, or purely elliptic groups (a torus generated by a single elliptic element).

work in infinite volume, where standard spectral methods (decomposition into Maass forms and Eisenstein series) cannot be applied. Therefore we develop novel "soft" methods using only operator theory (the abstract spectral theorem; see e.g. [Hal63]) and spectral information from Patterson-Sullivan theory [Pat76, Sul84] and Lax-Phillips [LP82] to circumvent explicit knowledge of a spectral decomposition.

Remark 1.2. Note that an archimedean count is also used in Theorem 2 of [BGS06]; Corollary 1.8 in the present paper is closely related to this result. The difference is counting in a group versus counting in an orbit. The problem of counting in the group in infinite volume was solved in Lax-Phillips [LP82]. To count in an orbit, one faces the serious issue of a stabilizer, as we discuss below.

To state our main theorem, we require some notation. Our function of choice will be the sum of two squares:

$$f(c, d) := c^2 + d^2. \tag{1.1}$$

Let $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ be a non-elementary finitely-generated Fuchsian group. We will soon turn our attention exclusively to groups of the second kind – ones having infinite co-volume – but we do not make this restriction just yet. Denote by Γ_∞ the set of elements of Γ which stabilize infinity:

$$\Gamma_\infty := \left\{ \gamma \in \Gamma \mid \gamma = \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \right\}. \tag{1.2}$$

Let $\delta = \delta(\Gamma) \leq 1$ denote the Hausdorff dimension of the limit set of Γ and $\theta < \delta$ be the spectral gap (see §2.2). The index of Γ in $\mathrm{SL}_2(\mathbb{Z})$ is finite if and only if $\delta = 1$.

Let $\mathcal{O}$ be the orbit of bottom rows of Γ,

$$\mathcal{O} := (0, 1)\Gamma = \left\{ (c, d) : \begin{pmatrix} * & * \\ c & d \end{pmatrix} \in \Gamma_\infty \backslash \Gamma \right\}, \tag{1.3}$$

and for a height $T > 1$ let $\mathcal{O}(T)$ denote the set of orbital points not exceeding this height:

$$\mathcal{O}(T) := \{ (c, d) \in \mathcal{O} \mid c^2 + d^2 < T \}. \tag{1.4}$$

For $R \geq 1$ let $\mathcal{O}(T, R)$ be defined by

$$\mathcal{O}(T, R) := \{ (c, d) \in \mathcal{O}(T) \mid f(c, d) \text{ has at most } R \text{ prime factors} \}.$$

Recall the notation

$$f \asymp g \quad \text{for} \quad g \ll f \ll g.$$

Theorem 1.3 (Main Theorem)**.** *Let $\Gamma \subset SL_2(\mathbb{Z})$ be a non-elementary finitely-generated Fuchsian group, let δ be the Hausdorff dimension of its limit set, θ the spectral gap, and $\mathcal{O}(T)$ and $\mathcal{O}(T, R)$ defined as above. Assume Γ_∞ is nontrivial (then $\delta > 1/2$, [Bea68]). As $T \to \infty$,*

(1) *There exist constants $c_0 > 0$ and $\eta > 0$ such that*

$$|\mathcal{O}(T)| = c_0 T^\delta + O(T^{\delta - \eta}). \tag{1.5}$$

(2) *For any fixed $R > 4/(\delta - \theta)$,*

$$|\mathcal{O}(T, R)| \asymp T^\delta / \log T. \tag{1.6}$$

Part (1) above tells us that the orbit is thin if and only if $\delta < 1$, and is the quintessence of our necessity to work in infinite co-volume. In [BGS06], it is proved that if $\delta > 1/2$, then there always exists *some* spectral gap $\theta < \delta$. Gamburd [Gam02] shows that if $\delta > 5/6$ then we can take $\theta = 5/6$. If on the other hand Γ is a finite co-volume group and moreover a congruence group, then Kim-Sarnak [KS03] allows $\theta = 1/2 + 7/64 \approx .609$. The following corollary is immediate.

Corollary 1.4. *As $T \to \infty$,*

$$|\mathcal{O}(T, R)| \gg T^\delta / \log T$$

for

(i) *$R = 25$ if $\delta > 149/150$ (setting $\theta = 5/6$),*
(ii) *$R = 11$ if Γ is a finite index congruence group (where $\delta = 1$ and $\theta = 39/64$), and*
(iii) *$R = 9$ if $\delta - \theta > 4/9$.*

It is known that in dimension two, infinite co-volume groups exist with δ arbitrarily close to 1 (e.g. [Gam02]) so part (i) above is not vacuous; moreover the example of Γ with δ arbitrarily close to 1 given in [Gam02] does in fact contain unipotent elements. Part (ii) is not particularly interesting, since one can vastly improve R for a finite index congruence group with classical techniques. We include it here only for comparison. It is not known whether part (iii) above is vacuous (even taking $\delta = 1$ requires $\theta < 5/9 \approx .555$ and is outside the reach of Kim-Sarnak), but $R = 9$ is the limit of our methods, and coincidentally is precisely the number of factors attained in Brun's original attack [Bru19] on the twin prime conjecture.

Remark 1.5. Determining membership in $\mathcal{O}$ amounts to expressing an element in Γ as a word in the generators – not an easy task. We are finding numbers with few prime factors despite having extremely limited knowledge as to which numbers appear!

Remark 1.6. The choice of the function $f(c, d) = c^2 + d^2$ is cosmetic; our methods apply to an arbitrary polynomial f, and we plan to detail this generalization in a future publication. That said, our current choice of f is natural, not only historically (indeed the problem of finding primes in sums of two squares dates back to Fermat) but also because, as we shall see, it is everywhere unobstructed in the affine linear sieve (akin to looking for primes congruent 1 modulo q – there are no obstructions for any q). See Remark A.2.

Remark 1.7. The requirement in the Main Theorem that Γ_∞ be nontrivial is undesirable. Indeed the conclusions should hold without this assumption, and we are currently working to remove it by other methods. See Remark B.2.

More importantly, unipotent elements furnish an affine injection into our orbit, enabling more classical sieve techniques. Precisely, if $\alpha, \gamma \in \Gamma$ and $\alpha \neq I$ fixes infinity, then $f\big((0, 1)\gamma\alpha^n\big)$ is a quadratic polynomial in n. These are known [Iwa78] to contain infinitely many numbers with at most two factors! Moreover by varying γ, one can accrue a Zariski dense set of 2-almost primes, in the sense of [BGS06]. Therefore we state one more immediate corollary which cannot be deduced by "cheating" with unipotents:

Corollary 1.8. *Let Γ, δ and $\mathcal{O}(T, R)$ be as in the Main Theorem. Then for $R = 1$ we have the following upper bound for the number of primes in f:*

$$|\mathcal{O}(T, 1)| \ll T^\delta / \log T, \qquad as\ T \to \infty.$$

This is off by a constant multiple from the expected asymptotic formula.

This paper is organized as follows. In the next section we give background material on Strong Approximation, the geometry and spectra of infinite co-volume groups, and the weighted linear β-sieve. In §3 we prove a certain Main Identity, which shows how to grow the lattice point count at time T from that at small times via the Laplace operator (this is the key to circumventing an explicit spectral theorem). In §4 we collect preliminary facts about infinite volume lattice point counts before proving the Main Theorem in §5. Some technical issues are reserved for the Appendices.

2. Background Material

2.1. **Strong Approximation.** Our first ingredients are algebraic in nature. We require the Strong Approximation Theorem of Matthews, Vaserstein, and Weisfeiler [MVW84], stating that if $L \subset \mathrm{SL}_n(\mathbb{Z})$ is Zariski dense in $\mathrm{SL}_n(\mathbb{Z})$ then the projection of L on $\mathrm{SL}_n(\mathbb{Z}_p)$ is dense for all but finitely many primes p. Recall that the ring of p-adic integers $\mathbb{Z}_p$ is the inverse limit of the finite rings $\mathbb{Z}/p^k\mathbb{Z}$. In particular, this means L is onto $\mathrm{SL}_n(\mathbb{Z}/p\mathbb{Z})$. Actually in dimension two this can be done by more elementary methods [DSV03].

We also require Goursat's Lemma which states the following. Let G_1, G_2 be groups, and let H be a subgroup of $G_1 \times G_2$ such that the two projections $p_j : H \to G_j$, $j = 1, 2$ are surjective. Let N_j be the kernel of p_j. One can identify N_1 as a normal subgroup of G_2, and N_2 as a normal subgroup of G_1. Then the image of H in $G_1/N_2 \times G_2/N_1$ is the graph of an isomorphism $G_1/N_2 \approx G_2/N_1$. As the group $\mathrm{PSL}_2(\mathbb{F})$ is simple for a finite field $\mathbb{F}$ satisfying $|\mathbb{F}| > 4$, this implies that if $p_1 \neq p_2$ are two primes and $L \subset \mathrm{SL}_2(\mathbb{Z})$ projects onto both $SL_2(\mathbb{Z}/p_1\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z}/p_2\mathbb{Z})$, then L projects onto $SL_2(\mathbb{Z}/p_1p_2\mathbb{Z})$.

We follow [BGS06] in recording these two facts as follows:

Theorem 2.1. *Let $\Gamma \subset SL_2(\mathbb{Z})$ be nonelementary. Then there exists a number $\mathfrak{B}$ which is the product of a finite set of "ramified" primes such that if $q = q'q''$ is square-free with $q'|\mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$ then the projection of Γ in $SL_2(\mathbb{Z}/q\mathbb{Z})$ is the product $G_{q'} \times SL_2(\mathbb{Z}/q''\mathbb{Z})$, where $G_{q'}$ is the projection of Γ in $SL_2(\mathbb{Z}/q'\mathbb{Z})$.*

2.2. **Infinite-volume geometry and spectral data.** Let $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ be a finitely-generated non-elementary Fuchsian group. Then Γ acts on the Poincaré upper half plane $\mathbb{H}$ by fractional linear transformations. In dimension two, being finitely generated is equivalent to being geometrically finite, i.e. that the Riemann surface $\mathcal{F} = \Gamma \backslash \mathbb{H}$ has finitely many bounding sides [Bea83].

As the action of Γ on $\mathbb{H}$ is discrete, there are no limit points in $\mathbb{H}$. There are however limit points in the boundary $\widehat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. The set of all limit points of Γ is called the limit set $\Lambda = \Lambda(\Gamma)$. It is a Cantor-like fractal and has some Hausdorff dimension $\delta = \delta(\Gamma) \in [0, 1]$. A geometrically finite group has $\delta < 1$ if and only if $\mathcal{F}$ has infinite hyperbolic volume (i.e. Γ has infinite index in $\mathrm{SL}_2(\mathbb{Z})$).

If we label the generators of $\mathrm{SL}_2(\mathbb{Z})$ by $T : z \mapsto z + 1$ and $S : z \mapsto -1/z$, then a prototypical example of the type of group we have in mind is the Hecke group $\Gamma = \langle T^4, S \rangle$. This group has fundamental domain $\mathcal{F} = \{z \in \mathbb{H} \mid |z| > 1, |\mathfrak{Re}(z)| < 2\}$, whose vertical
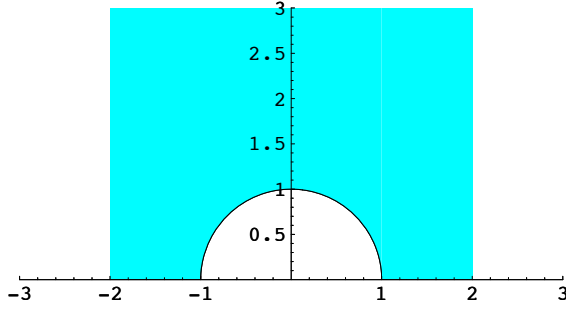
FIGURE 1. A fundamental domain for $\Gamma = \langle T^4, S \rangle$.

strips touch the real line at a free boundary and clearly contribute infinite hyperbolic volume. See Fig. 1. We will assume throughout that $\mathcal{F}$ has a cusp at infinity, as in this example.

As usual, $\mathcal{F}$ is equipped with a hyperbolic volume element $dz = \frac{dx\,dy}{y^2}$. The positive-definite Laplace-Beltrami operator $\Delta = -y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$ of $\mathcal{F}$ acts in the space of C^∞ functions with compact support $C_K^\infty(\mathcal{F})$, and has a unique self-adjoint extension to an unbounded operator on $L^2(\mathcal{F})$. Denote the spectrum of Δ on $L^2(\Gamma \backslash \mathbb{H})$ by $\mathrm{Spec}(\Gamma \backslash \mathbb{H})$.

The spectrum of Δ below $1/4$ consists only of finitely-many point eigenvalues and the tempered spectrum contained in $[1/4, \infty)$ is purely continuous [LP82]. Notice that in finite volume $\delta = 1$ and the base eigenvalue is $\lambda_0 = 0$, corresponding to the constant function $\phi_0 = 1/\sqrt{\mathrm{vol}(\mathcal{F})}$, scaled to have unit L^2-norm. Neither does this make sense if $\mathrm{vol}(\mathcal{F}) = \infty$ nor is any non-zero constant function square-integrable in this case. It follows from the work of Patterson [Pat76, Sul84] that $\delta > 1/2$ if and only if

$$\mathrm{Spec}(\Gamma \backslash \mathbb{H}) \cap [0, 1/4) \neq \varnothing,$$

in which case $\lambda_0 = \delta(1-\delta)$ is the base eigenvalue of Δ. This eigenvalue is isolated, has multiplicity one, and any associated eigenfunction is of constant sign on $\mathcal{F}$; in particular we can choose it to be nonnegative. The rôle of the constant function (volume) is then played by the base eigenfunction ϕ_0 which Patterson determined explicitly as the integral of a Poisson kernel against the so-called Patterson-Sullivan measure μ, supported on the limit set Λ:

$$\phi_0(x, y) := \int_\Lambda \left(\frac{(t^2 + 1)y}{(x - t)^2 + y^2} \right)^\delta d\mu(t). \qquad (2.1)$$

If $\mathcal{F}$ has a cusp (which we assume it does) then $\delta > 1/2$.

Let $\Gamma(q)$ denote the principal "congruence" subgroup of Γ of level q,

$$\Gamma(q) := \{\gamma \in \Gamma : \gamma \equiv I(\mathrm{mod}\, q)\}. \tag{2.2}$$

This is of course still a Fuchsian group of the second kind, but has finite index in Γ (and therefore also has the same limit set and Hausdorff dimension – every Cauchy sequence in $\mathbb{H}$ under the action of Γ has a corresponding sequence under $\Gamma(q)$ with the same limit point). The inclusion $\Gamma(q) \subset \Gamma$ induces the reverse inclusion

$$\mathrm{Spec}(\Gamma\backslash\mathbb{H}) \subset \mathrm{Spec}(\Gamma(q)\backslash\mathbb{H}).$$

In particular this means the base eigenfunction ϕ_0 is an "oldform" on $\Gamma(q)$ (but must be rescaled to have unit L^2-norm).

Corresponding to any point eigenvalue $\lambda \in \mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})$ is an eigenfunction, $\phi \in L^2(\Gamma(q)\backslash\mathbb{H})$. It may be the case that ϕ (which a priori is only $\Gamma(q)$-automorphic) is also automorphic with respect to Γ. In this case we we call ϕ an "oldform" and λ is an "oldvalue". In the opposite case we call ϕ a "newform" and λ a "newvalue". Denote by $\mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})_{new}$ the subset of $\mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})$ consisting of "new" eigenvalues.

Definition 2.2. *We say that Γ has a* **spectral gap** *$\theta \in [1/2, \delta)$ if there exists $\mathfrak{B} \in \mathbb{N}$ such that for $q = q'q''$ square-free with $q'|\mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$, we have*

$$\mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})_{new} \cap (0, \theta(1-\theta)) \subset \mathrm{Spec}(\Gamma(q')\backslash\mathbb{H})_{new}. \tag{2.3}$$

Note that this definition of the "spectral gap" is not the conventional one (for which see, for example, [EMV07, KS08]).

Collecting the results in [BG07, BGS06] and their extension from prime to square-free of [Gam02] we have:

Theorem 2.3 ([Gam02, BG07, BGS06]). *Let $\Gamma \subset SL_2(\mathbb{Z})$ and $\delta > 1/2$ be the Hausdorff measure of its limit set.*

    (1) *There exists a spectral gap $\theta \in [1/2, \delta)$ such that (2.3) holds with $\mathfrak{B} = 1$.*

    (2) *If $\delta > 5/6$ then (2.3) holds with $\theta = 5/6$ and the number $\mathfrak{B}$ is precisely the one which appeared in Theorem 2.1.*

The case $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, $\mathfrak{B} = 1$ and $\theta = 1/2$ is the celebrated (and unsolved) Selberg 1/4-Conjecture [Sel65], which in modern parlance is a consequence of the generalized Ramanujan Conjectures.

We now record the abstract spectral theorem for unbounded self-adjoint operators as follows.

Theorem 2.4 (Abstract Spectral Theorem). *There exists a spectral measure ν, supported on $\mathrm{Spec}(\Gamma\backslash\mathbb{H})$ and a unitary spectral operator $\widehat{}: L^2(\Gamma\backslash\mathbb{H}) \to L^2((0,\infty), d\nu)$ such that:*

(1) *We have the Abstract Parseval's Identity: for $\phi_1, \phi_2 \in L^2(\Gamma\backslash\mathbb{H})$,*

$$\langle \phi_1, \phi_2 \rangle_{L^2(\Gamma\backslash\mathbb{H}, dz)} = \left\langle \widehat{\phi_1}, \widehat{\phi_2} \right\rangle_{L^2(\mathrm{Spec}(\Gamma\backslash\mathbb{H}), d\nu)} \tag{2.4}$$

(2) *The spectral operator $\widehat{}$ is diagonal with respect to the Laplacian: for $\phi \in L^2(\Gamma\backslash\mathbb{H})$ and $\lambda \in \mathrm{Spec}(\Gamma\backslash\mathbb{H})$*

$$\widehat{\Delta\phi}(\lambda) = \lambda\, \widehat{\phi}(\lambda). \tag{2.5}$$

(3) *If λ is a point eigenvalue of multiplicity one with associated L^2 eigenfunction ϕ_λ of unit norm, then for any $\phi \in L^2(\Gamma\backslash\mathbb{H})$,*

$$\widehat{\phi}(\lambda) = \langle \phi, \phi_\lambda \rangle. \tag{2.6}$$

2.3. **Combinatorial Sieve.** Let $\mathcal{A}$ be a sequence of non-negative real numbers $\{a_n\}_n$ of which all but finitely many are zero. For $R \geq 1$ let Z_R denote the set of positive integers with at most R prime divisors. The main objective in sieve theory is to determine lower bound estimates for

$$\sum_{n \in Z_R} a_n \tag{2.7}$$

given knowledge of how $\mathcal{A}$ is distributed along each of the arithmetic progressions $0 (\mathrm{mod}\, q)$ for square-free q. In the following setup of the sieve, there are many parameters. Their heuristic meaning is as follows.

Let $q \geq 1$ be square-free and collect the elements of $\mathcal{A}$ whose index is divisible by q via

$$\mathcal{A}_q := \{a_n \in \mathcal{A} \mid n \equiv 0(q)\}. \tag{2.8}$$

Of course $\mathcal{A}_1 = \mathcal{A}$. The parameter T is the cut-off point after which all a_n's are zero. The parameter $\mathfrak{B} \in \mathbb{N}$ is the product of a fixed finite set of "ramified" primes. We will decompose q into the $\mathfrak{B}$-part and the rest: $q = q'q''$ with $q'|\mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$. The parameter $\mathcal{X}$ is an approximation to $|\mathcal{A}| := \sum_n a_n$, and $\omega : \mathbb{N} \to [0, 1]$ represents the local density at q. Thus $\omega(q)\mathcal{X}$ is an approximation to

$$|\mathcal{A}_q| = \sum_{n \equiv 0(q)} a_n.$$

At the $\mathfrak{B}$-part, there is a lower order term, $\mathcal{X}_{q'}$ such that

$$r(q) := |\mathcal{A}_q| - \omega(q)(\mathcal{X} + \mathcal{X}_{q'})$$

is an error term. The error is small on average: there is a sieving level Q such that the total contribution from the error terms up to Q is a power savings off the main term $\mathcal{X}$.

Precisely, we require the following conditions:

(S1) $\mathcal{X} \asymp |\mathcal{A}| := \sum_n a_n$ and $\mathfrak{B} \in \mathbb{N}$ is a fixed natural number. There is a parameter $T > 1$ such that

$$a_n = 0 \text{ if } n > T. \tag{2.9}$$

(S2) The function $\omega : N \to [0, 1]$ satisfies $\omega(1) = 1$ and $\omega(q) < 1$ for $q > 1$. Moreover, ω is multiplicative away from $\mathfrak{B}$. By this we mean that for any q square-free, write $q = q'q''$ with $q'|\mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$. Then $\omega(q) = \omega(q') \prod_{p|q''} \omega(p)$.

(S3) There exists a fixed constant $K < \infty$ such that for $2 \le v \le z$, we have the following local density bound:
$$\prod_{\substack{v \le p \le z \\ p \nmid \mathfrak{B}}} (1 - \omega(p))^{-1} \le \left(\frac{\log z}{\log v} \right) \left(1 + \frac{K}{\log v} \right).$$

(S4) For any divisor $q'|\mathfrak{B}$ let $\mathcal{X}_{q'}$ satisfy $\mathcal{X}_{q'} \ll \mathcal{X}^{1-\eta}$ for some $\eta > 0$.

(S5) Define $r(q) := |\mathcal{A}_q| - \omega(q)(\mathcal{X} + \mathcal{X}_{q'})$ and assume that for some parameter $Q > 1$ and $\epsilon > 0$,

$$\sum_{q \le Q} |r(q)| \ll \mathcal{X}^{1-\epsilon}.$$

The following Theorem is a simple consequence of Theorem 3.7 on page 63 in [Iwa96]. We derive it from the original in Appendix A.

Theorem 2.5. *Let $\mathcal{A}$ be as described above. Then*

$$\sum_{n \in Z_R} a_n \asymp \mathcal{X} \prod_{\substack{p \le Q \\ p \nmid \mathfrak{B}}} (1 - \omega(p)), \tag{2.10}$$

for any R satisfying

$$R > 2 \log T / \log Q. \tag{2.11}$$

In our application, Goursat's Lemma demonstrates the multiplicativity of the local density ω in (S2), and Strong Approximation allows us to compute ω explicitly and verify (S3). The finite collection of primes comprising $\mathfrak{B}$ which may have exceptional eigenvalues in Theorem 2.3 are directly responsible for the extra terms $\mathcal{X}_{q'}$ in (S4) but their contribution is harmless to the inclusion-exclusion of the sieve.

3. The Main Identity

Let $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$. Recall the classical fact that for $\gamma = \begin{pmatrix} * & * \\ c & d \end{pmatrix} \in \Gamma$ we have

$$\mathfrak{Im}(\gamma z) = \frac{\mathfrak{Im} z}{|cz + d|^2}.$$

Fix $T > 1$ and define the characteristic function

$$\chi_T(z) := \begin{cases} 1 & \text{if } \mathfrak{Im}(z) > 1/T \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

so that

$$\chi_T(\gamma\mathrm{i}) = \begin{cases} 1 & \text{if } c^2 + d^2 < T \\ 0 & \text{otherwise.} \end{cases}$$

Let $\Gamma_\infty \subset \Gamma$ be the maximal subgroup stabilizing infinity; clearly χ_T is Γ_∞-invariant.

We average χ_T over the group Γ

$$F_T(z) := \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \chi_T(\gamma z), \tag{3.2}$$

so that we can recover the lattice point count (1.4) via

$$F_T(\mathrm{i}) = |\mathcal{O}(T)|.$$

Clearly F_T is Γ-invariant, so is well defined as a function on $\Gamma \backslash \mathbb{H}$.

Lemma 3.1. $F_T \in L^2(\Gamma \backslash \mathbb{H})$ *if and only if* Γ_∞ *is nontrivial.*

Proof. If Γ_∞ is nontrivial, let N be the finite interval corresponding to the width of the cusp at infinity. Otherwise set $N = \mathbb{R}$. By unfolding we have

$$\begin{aligned} \|F_T\|^2 &= \int_{\Gamma \backslash \mathbb{H}} F_T(z) F_T(z) dz = \int_{\Gamma \backslash \mathbb{H}} \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \chi_T(\gamma z) F_T(z) dz \\ &= \int_{\Gamma_\infty \backslash \mathbb{H}} \chi_T(z) F_T(z) dz = \int_{1/T}^{\infty} \int_N F_T(z) dx \frac{dy}{y^2}. \end{aligned}$$

For $\mathfrak{Im}(z)$ sufficiently large, only the term $\gamma = I$ contributes to the sum (3.2), and so $F_T(z) = 1$. It is now clear that the above integral converges if and only if $N \neq \mathbb{R}$. $\square$

Instead of accessing F_T directly, we will seek an identity which "grows" the count at time T from the count at time $T = 1$ (and at some other time, $T = b$). Let $\phi \in L^2(\Gamma \backslash \mathbb{H})$ be an eigenfunction of the hyperbolic Laplace-Beltrami operator $\Delta = -y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$ with eigenvalue

$\lambda = s(1-s) < 1/4$. By unfolding the following inner product and using the fact that the constant (or any) Fourier coefficient of ϕ satisfies the same differential equation as ϕ itself, we have

$$
\begin{aligned}
\langle F_T, \phi \rangle &= \int_{\Gamma \backslash \mathbb{H}} F_T(z) \bar{\phi}(z) dz = \int_{\Gamma_\infty \backslash \mathbb{H}} \chi_T(z) \bar{\phi}(z) dz \\
&= \int_{1/T}^\infty \left(\int_N \bar{\phi}(z) dx \right) \frac{dy}{y^2} = \int_{1/T}^\infty (\alpha y^s + \beta y^{1-s}) \frac{dy}{y^2} \\
&= A_\phi T^s + B_\phi T^{1-s},
\end{aligned}
$$

where α, β, A_ϕ, and B_ϕ are some constants depending on the eigenfunction ϕ. (As F_T is like a truncated Eisenstein series, this calculation is just the incomplete Mellin transform of the constant term of ϕ.)

Following the methodology of Selberg [Sel56], we seek an identity which depends only on the eigenvalue λ or s but not on the eigenfunction ϕ. Reformulate the above identity as a dot product of vectors

$$
\langle F_T, \phi \rangle = A_\phi T^s + B_\phi T^{1-s} = (T^s, T^{1-s}) \begin{pmatrix} A_\phi \\ B_\phi \end{pmatrix} \tag{3.3}
$$

and write it at time $T = 1$ and some other time $T = b > 1$:

$$
\begin{aligned}
\langle F_1, \phi \rangle &= A_\phi + B_\phi, \\
\langle F_b, \phi \rangle &= A_\phi b^s + B_\phi b^{1-s},
\end{aligned}
$$

or

$$
\begin{pmatrix} \langle F_1, \phi \rangle \\ \langle F_b, \phi \rangle \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ b^s & b^{1-s} \end{pmatrix} \begin{pmatrix} A_\phi \\ B_\phi \end{pmatrix}.
$$

Notice that the matrix on the right hand side only depends on the eigenvalue s. We multiply both sides by the inverse matrix

$$
\begin{pmatrix} A_\phi \\ B_\phi \end{pmatrix} = \frac{1}{b^{1-s} - b^s} \begin{pmatrix} b^{1-s} & -1 \\ -b^s & 1 \end{pmatrix} \begin{pmatrix} \langle F_1, \phi \rangle \\ \langle F_b, \phi \rangle \end{pmatrix},
$$

and insert this into (3.3):

$$
\begin{aligned}
\langle F_T, \phi \rangle &= (T^s, T^{1-s}) \frac{1}{b^{1-s} - b^s} \begin{pmatrix} b^{1-s} & -1 \\ -b^s & 1 \end{pmatrix} \begin{pmatrix} \langle F_1, \phi \rangle \\ \langle F_b, \phi \rangle \end{pmatrix} \\
&= K_T(s) \langle F_1, \phi \rangle + L_T(s) \langle F_b, \phi \rangle, \tag{3.4}
\end{aligned}
$$

where

$$
\begin{aligned}
K_T(s) &= \frac{T^s b^{1-s} - T^{1-s} b^s}{b^{1-s} - b^s}, \tag{3.5} \\
L_T(s) &= \frac{T^{1-s} - T^s}{b^{1-s} - b^s}
\end{aligned}
$$

are functions which only depend on the eigenvalue, $\lambda = s(1-s)$.

Notice that for $T > 1$, $s \in (1/2, 1]$, and b fixed

$$K_T(s), L_T(s) \ll T^s. \tag{3.6}$$

For $s = \frac{1}{2} + it$ we have

$$K_T(s) = T^{1/2} \frac{\sin(t \log T / \log b)}{\sin(t \log b)},$$

$$L_T(s) = \left(\frac{T}{b}\right)^{1/2} \frac{\sin(t \log T)}{\sin(t \log b)}.$$

Choosing $b = b(T)$ such that $\frac{\log T}{\log b} \in \mathbb{Z}$ ensures that the functions above are

$$\ll T^{1/2} \log T.$$

For example, we can take

$$b = \exp\left(\frac{\log T}{\lceil \log T \rceil}\right) \in (1, e], \tag{3.7}$$

where $\lceil \cdot \rceil$ is the ceiling function, returning the smallest integer not less than its argument. In particular,

$$b < 3.$$

As $\lambda = s(1-s)$, we abuse notation by writing

$$K_T(\lambda) \text{ and } L_T(\lambda)$$

in place of

$$K_T(s) \text{ and } L_T(s).$$

This should cause no confusion. Just as one can exponentiate a matrix, one can define a function of a differential operator, which is itself a differential operator. So as ϕ is an eigenfunction satisfying $\Delta \phi = \lambda \phi$, we have

$$K_T(\Delta)\phi = K_T(\lambda)\phi,$$

where the left hand side acts by differentiation and the right hand side is multiplication by a function of the eigenvalue. The same holds for L_T.

Since the Laplacian Δ is self-adjoint, we have from (3.4)

$$
\begin{aligned}
\langle F_T, \phi \rangle &= K_T(\lambda) \langle F_1, \phi \rangle + L_T(\lambda) \langle F_b, \phi \rangle \\
&= \langle F_1, K_T(\lambda)\phi \rangle + \langle F_b, L_T(\lambda)\phi \rangle \\
&= \langle F_1, K_T(\Delta)\phi \rangle + \langle F_b, L_T(\Delta)\phi \rangle \\
&= \langle K_T(\Delta)F_1, \phi \rangle + \langle L_T(\Delta)F_b, \phi \rangle \\
&= \langle K_T(\Delta)F_1 + L_T(\Delta)F_b, \phi \rangle .
\end{aligned}
$$

This identity holds for any eigenfunction ϕ in the point spectrum, and so it should be the case that

$$F_T(z) = K_T(\Delta)F_1(z) + L_T(\Delta)F_b(z) \tag{3.8}$$

holds in general. Notice that the T dependence has been entirely removed from F on the right hand side and only appears in the spectral operators K_T and L_T! It should now be clear what we mean by gathering information about F at time T using only the Laplacian and F at small times. We reserve the rest of this section for the proof of this Main Identity.

Theorem 3.2 (Main Identity). *For fixed $T \geq 1$ let F_T be defined by (3.2). Then there exists a number $b < 3$ and functions K_T and L_T satisfying:*

$$L_T(\lambda) \ll K_T(\lambda) \ll T^s$$

for $\lambda = s(1-s)$ and $s \in (1/2, 1]$, and

$$L_T(\lambda) \ll K_T(\lambda) \ll T^{1/2}\log T$$

for $\lambda \geq 1/4$, such that (3.8) holds for almost every z.

The argument above (3.8) proves this identity along the point spectrum. If we had an explicit spectral theorem in infinite volume, we would just need to carry out similar computations to prove this identity on the continuous spectrum (for a finite co-volume group, this is easily achieved via the Eisenstein series). Instead, we will use ideas from "almost" eigenfunctions and perturbation theory for the proof, which will occupy the remainder of this section.

We begin with a technical lemma. Note that from now on, ϕ is not assumed to be an eigenfunction of Δ, only that $\phi \in L^2(\Gamma \backslash \mathbb{H})$.

Lemma 3.3. *For any $\phi \in L^2(\Gamma \backslash \mathbb{H})$ and any $\lambda = s(1-s) \geq 0$, $\lambda \neq 1/4$ there exist constants A and B such that*

$$\langle F_T, \phi \rangle = A T^s + B T^{1-s} + O_{\lambda,T,\Gamma}\left(\|(\Delta - \lambda)\phi\|\right). \tag{3.9}$$

If $\lambda = 1/4$ then

$$\langle F_T, \phi \rangle = A T^{1/2} + B T^{1/2}\log T + O_{\lambda,T,\Gamma}\left(\|(\Delta - \lambda)\phi\|\right).$$

This is an explicit computation, applying the method of variation of parameters and estimating the inhomogeneous component (which relies on Γ_∞ being non-trivial). We present the details in Appendix B.

Remark 3.4. One should think of ϕ as an "almost" eigenfunction with eigenvalue λ. Then the error term in (3.9) should be small. This is a heuristic only; the argument applies for general ϕ and λ, and we will soon see that such an application is necessary.

The following proposition says that the difference of the right and left hand sides of (3.9) has no correlation with any almost eigenfunction.

Proposition 3.5. *Let T, b, and F_T be as above and set*

$$G_T = F_T - K_T(\Delta)F_1 - L_T(\Delta)F_b. \qquad (3.10)$$

Then for any $\phi \in L^2(\Gamma\backslash\mathbb{H})$ and any $\lambda \geq 0$ we have

$$\langle G_T, \phi \rangle \ll_{\lambda, T} \; \|(\Delta - \lambda)\phi\|. \qquad (3.11)$$

Proof. Fix an arbitrary $\phi \in L^2(\Gamma\backslash\mathbb{H})$ and $\lambda = s(1-s) \geq 0$, and let G_T be defined by (3.10). Assume $\lambda \neq 1/4$ (the computation in the the case of $\lambda = \frac{1}{4}$ is similar). Consider the following trivial identity obtained by adding and subtracting identical terms:

$$
\begin{aligned}
\langle G_T, \phi \rangle \;=\; & \langle F_T, \phi \rangle - (A_\phi T^s + B_\phi T^{1-s}) \\
& - (\langle K_T(\Delta)F_1, \phi \rangle - K_T(\lambda)\langle F_1, \phi \rangle) \\
& - (\langle L_T(\Delta)F_b, \phi \rangle - L_T(\lambda)\langle F_b, \phi \rangle) \\
& - (K_T(\lambda)\langle F_1, \phi \rangle - K_T(\lambda)(A_\phi + B_\phi)) \\
& - \left(L_T(\lambda)\langle F_b, \phi \rangle - L_T(\lambda)(A_\phi b^s + B_\phi b^{1-s})\right) \\
& + (A_\phi T^s + B_\phi T^{1-s}) - K_T(\lambda)(A_\phi + B_\phi) - L_T(\lambda)(A_\phi b^s + B_\phi b^{1-s}).
\end{aligned}
\qquad (3.12)
$$

The bottom row is zero by construction of K_T and L_T in (3.5). The top row is $\ll_{\lambda, T} \|(\Delta - \lambda)\phi\|$ by Lemma 3.3, as are the fourth and fifth rows. It remains to understand the second and third rows.

Let $\widehat{\phi}$ denote the spectral transform of ϕ and let $\widehat{F_T}$ be the spectral transform of F_T. As λ is fixed, let λ' be in $\mathrm{Spec}(\Gamma\backslash\mathbb{H})$. By the Mean Value Theorem and Cauchy-Schwarz, the second row is

$$
\begin{aligned}
(3.12) \;=\;\; & \langle K_T(\Delta)F_1, \phi \rangle - K_T(\lambda)\langle F_1, \phi \rangle \\
=\;\; & \left\langle K_T \cdot \widehat{F_1}, \widehat{\phi} \right\rangle - K_T(\lambda)\left\langle \widehat{F_1}, \widehat{\phi} \right\rangle \\
=\;\; & \int_{\mathrm{Spec}(\Gamma\backslash\mathbb{H})} (K_T(\lambda') - K_T(\lambda))\widehat{F_1}(\lambda')\overline{\widehat{\phi}}(\lambda')d\nu(\lambda') \\
\ll_{\lambda, T}\;\; & \int_{\mathrm{Spec}(\Gamma\backslash\mathbb{H})} (\lambda' - \lambda)\overline{\widehat{\phi}}(\lambda')\widehat{F_1}(\lambda')d\nu(\lambda') \\
\ll\;\; & \|(\Delta - \lambda)\phi\|\|F_1\|.
\end{aligned}
$$

The calculation for the third row is identical and we are done.

$\square$

Finally, we show that a function which is uncorrelated to any almost eigenfunction is zero.

Proof of the Main Identity. We aim to show that G_T defined in (3.10) vanishes almost everywhere.

Since (3.11) holds for any ϕ, we are free to choose our ϕ. Equivalently we may choose its spectral transform $\widehat{\phi}$, so we make the following construction. Fix $\varepsilon > 0$ and fix an arbitrary $\lambda > 0$. Let

$$\widehat{\phi}(\lambda') := \begin{cases} \widehat{G}_T(\lambda') & \text{if } \lambda' \in (\lambda - \varepsilon, \lambda + \varepsilon) \\ 0 & \text{otherwise,} \end{cases} \qquad (3.13)$$

where $\widehat{G}_T$ is the spectral transform of G_T. Inserting (3.13) into Abstract Parseval's Theorem (2.4) we have:

$$\begin{aligned} \langle G_T, \phi \rangle &= \left\langle \widehat{G}_T, \widehat{\phi} \right\rangle = \int_{\text{Spec}(\Gamma \backslash \mathbb{H})} \widehat{G}_T(\lambda') \overline{\widehat{\phi}}(\lambda') d\nu(\lambda') \\ &= \int_{\lambda - \varepsilon}^{\lambda + \varepsilon} |\widehat{G}_T(\lambda')|^2 d\nu(\lambda'), \end{aligned} \qquad (3.14)$$

and

$$\begin{aligned} \|(\Delta - \lambda)\phi\| &= \left(\int_{\text{Spec}(\Gamma \backslash \mathbb{H})} \left| (\lambda' - \lambda)\widehat{\phi}(\lambda') \right|^2 d\nu(\lambda') \right)^{1/2} \\ &= \left(\int_{\lambda - \varepsilon}^{\lambda + \varepsilon} |\lambda' - \lambda|^2 \left| \widehat{G}_T(\lambda') \right|^2 d\nu(\lambda') \right)^{1/2} \\ &\leq \varepsilon \left(\int_{\lambda - \varepsilon}^{\lambda + \varepsilon} \left| \widehat{G}_T(\lambda') \right|^2 d\nu(\lambda') \right)^{1/2}. \end{aligned} \qquad (3.15)$$

Inserting (3.14) and (3.15) into (3.11), we have:

$$\int_{\lambda - \varepsilon}^{\lambda + \varepsilon} |\widehat{G}_T(\lambda')|^2 d\nu(\lambda') \ll_{\lambda, T} \varepsilon \left(\int_{\lambda - \varepsilon}^{\lambda + \varepsilon} |\widehat{G}_T(\lambda')|^2 d\nu(\lambda') \right)^{1/2}.$$

If the left side is zero, we are done. If not, we have for an arbitrary λ:

$$\int_{\lambda - \varepsilon}^{\lambda + \varepsilon} |\widehat{G}_T(\lambda')|^2 d\nu(\lambda') \ll_{\lambda, T} \varepsilon^2.$$

Let $f(\lambda) = \int_{\lambda' < \lambda} |\widehat{G}_T(\lambda')|^2 d\nu(\lambda')$. Then f is everywhere continuously differentiable with

$$f'(\lambda) = \lim_{\varepsilon \to 0} \frac{f(\lambda + \varepsilon) - f(\lambda - \varepsilon)}{2\varepsilon} = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{\lambda - \varepsilon}^{\lambda + \varepsilon} |\widehat{G}_T(\lambda')|^2 d\nu(\lambda') = 0.$$

So as $f' = 0$ uniformly and $f(0) = 0$, we have shown that $f \equiv 0$, and $G_T = 0$ a.e.

This concludes the proof of the Main Identity. $\qquad\qquad \square$

4. Preliminaries

4.1. **Sums over w_T.** Once again let $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ and let $\Gamma_\infty \subset \Gamma$ be the stabilizer of infinity. Fix $\varepsilon > 0$ and let

$$\psi = \psi_\varepsilon \in L^2(\Gamma_\infty \backslash \mathbb{H}) \tag{4.1}$$

be an ε-approximation to the identity about $z_0 = i$. By this we ask that ψ

(1) be smooth, nonnegative,
(2) have total mass one $\int_{\Gamma_\infty \backslash \mathbb{H}} \psi = 1$, and
(3) be supported in a small neighborhood about $z_0 = i$ (a ball of radius $\varepsilon/10$ will suffice).

Recall the characteristic function χ_T from (3.1). Define the function $w_T = w_{T,\varepsilon} : \Gamma_\infty \backslash \Gamma \to [0,1]$ by

$$w_T(\gamma) := \int_{\Gamma_\infty \backslash \mathbb{H}} \chi_T(z) \psi_\varepsilon(\gamma z) dz. \tag{4.2}$$

For $\gamma \in \Gamma$ having bottom row (c,d) it is easy to see that

$$w_{T,\varepsilon}(\gamma) = \begin{cases} 1 & \text{if } c^2 + d^2 < T/(1+\varepsilon) \\ 0 & \text{if } c^2 + d^2 > T/(1-\varepsilon). \end{cases} \tag{4.3}$$

We first prove the following lemma. Let Ξ be a subgroup of Γ having finite index in Γ. Then both groups Ξ and Γ have the same limit set, Hausdorff dimension δ and base eigenvalue $\lambda_0 = \delta(1-\delta)$. Assume further that Ξ contains Γ_∞.

Lemma 4.1. *Let $\Gamma_\infty \subset \Xi \subset \Gamma$ with $[\Gamma : \Xi] < \infty$ and w_T be defined by (4.2). Let $\lambda_1 = s_1(1 - s_1)$ be the first eigenvalue above the base in $\mathrm{Spec}(G \backslash \mathbb{H})$, and assume $\lambda_1 < 1/4$. Then for any fixed $\gamma \in \Gamma$, there exists a constant $c_{\Gamma,\varepsilon} > 0$ depending only on Γ and ε such that*

$$\sum_{\xi \in \Gamma_\infty \backslash \Xi} w_T(\xi\gamma) = \frac{c_{\Gamma,\varepsilon} T^\delta}{[\Gamma : \Xi]} + O\left(\frac{1}{\varepsilon} T^{s_1}\right), \tag{4.4}$$

where the implied constant depends on Γ but not Ξ, γ or ε. Moreover,

$$c_{\Gamma,\varepsilon} = c_\Gamma(1 + O(\varepsilon)), \tag{4.5}$$

where $c_\Gamma > 0$ does not depend on ε.
If $\lambda_1 \geq 1/4$, replace the error term in (4.4) by $O(\frac{1}{\varepsilon} T^{1/2} \log T)$.

Proof. Throughout we will suppress dependence on ε until it is convenient.

Let F_T^Ξ be the function defined on $\Xi\backslash\mathbb{H}$ which is the average of the characteristic function χ_T over the group Ξ:

$$F_T^\Xi(z) := \sum_{\xi\in\Gamma_\infty\backslash\Xi} \chi_T(\xi z). \tag{4.6}$$

Similarly average ψ over Ξ to get

$$\Psi_{\varepsilon,z_0}^\Xi(z) := \sum_{\xi\in\Gamma_\infty\backslash\Xi} \psi_{\varepsilon,z_0}(\xi z), \tag{4.7}$$

which is an ε-approximation to the identity about $z_0 = \mathrm{i}$ in $L^2(\Xi\backslash\mathbb{H})$. Notice that $\psi_{z_0}(\gamma^{-1}z) = \psi_{\gamma z_0}(z)$, where the latter function is an approximation to the identity about γz_0.

Replace γ in (4.4) by γ^{-1} for convenience. Input the definition (4.2) into the left hand side of (4.4) and repeatedly unfold and refold the integrals:

$$
\begin{aligned}
\sum_{\xi\in\Gamma_\infty\backslash\Xi} w_T(\xi\gamma^{-1}) &= \sum_{\xi\in\Gamma_\infty\backslash\Xi} \left(\int_{\Gamma_\infty\backslash\mathbb{H}} \chi_T(z)\psi_{z_0}(\xi\gamma^{-1}z)dz \right) \\
&= \sum_{\xi\in\Gamma_\infty\backslash\Xi} \left(\int_{\Gamma_\infty\backslash\mathbb{H}} \chi_T(z)\psi_{\gamma z_0}(\xi z)dz \right) \\
&= \int_{\Gamma_\infty\backslash\mathbb{H}} \chi_T(z) \left(\sum_{\xi\in\Gamma_\infty\backslash\Xi} \psi_{\gamma z_0}(\xi z) \right) dz \\
&= \int_{\Gamma_\infty\backslash\mathbb{H}} \chi_T(z) \Psi_{\gamma z_0}^\Xi(z)dz \\
&= \sum_{\xi\in\Gamma_\infty\backslash\Xi} \left(\int_{\Xi\backslash\mathbb{H}} \chi_T(\xi z)\Psi_{\gamma z_0}^\Xi(z)dz \right) \\
&= \int_{\Xi\backslash\mathbb{H}} F_T^\Xi(z)\Psi_{\gamma z_0}^\Xi(z)dz \\
&= \left\langle F_T^\Xi, \Psi_{\varepsilon,\gamma z_0}^\Xi \right\rangle_\Xi, \tag{4.8}
\end{aligned}
$$

where the inner product above $\langle\cdot,\cdot\rangle_\Xi$ is with respect to the Hilbert space $L^2(\Xi\backslash\mathbb{H})$. The above exchanges of summation and integration are justified since everything in sight is nonnegative and convergent – the sum on the left hand side of (4.8) has finitely many terms by (4.3).

By Abstract Parseval's Theorem (2.4),

$$\left\langle F_T^\Xi, \Psi_{\gamma z_0}^\Xi \right\rangle_\Xi = \left\langle \widehat{F}_T^\Xi, \widehat{\Psi}_{\gamma z_0}^\Xi \right\rangle_{\mathrm{Spec}(\Xi\backslash\mathbb{H})} \qquad (4.9)$$

$$= \widehat{F}_T^\Xi(\lambda_0)\widehat{\Psi}_{\gamma z_0}^\Xi(\lambda_0) + \int_{\mathrm{Spec}(\Xi\backslash\mathbb{H})-\{\lambda_0\}} \widehat{F}_T^\Xi(\lambda)\widehat{\Psi}_{\gamma z_0}^\Xi(\lambda)d\nu_\Xi(\lambda),$$

where ν_Ξ is the spectral measure on $\mathrm{Spec}(\Xi\backslash\mathbb{H})$ (see Theorem 2.4). By (2.6) and multiplicity one of the base eigenvalue $\lambda_0 = \delta(1-\delta)$ we have

$$\widehat{F}_T^\Xi(\lambda_0) = \left\langle F_T^\Xi, \phi_0^\Xi \right\rangle_\Xi, \qquad (4.10)$$

$$\widehat{\Psi}_{\gamma z_0}^\Xi(\lambda_0) = \left\langle \Psi_{\gamma z_0}^\Xi, \phi_0^\Xi \right\rangle_\Xi, \qquad (4.11)$$

where ϕ_0^Ξ is the Patterson-Sullivan base eigenfunction in $L^2(\Xi\backslash\mathbb{H})$, normalized to have unit norm. Recall this function is real and nonnegative. It is elementary to verify that

$$\phi_0^\Xi = \frac{1}{\sqrt{[\Gamma : \Xi]}}\phi_0^\Gamma. \qquad (4.12)$$

Unfolding (4.10) and inserting (4.12), we have

$$\widehat{F}_T^\Xi(\lambda_0) = \left\langle F_T^\Xi, \phi_0^\Xi \right\rangle_\Xi$$

$$= \int_{\Xi\backslash\mathbb{H}} F_T^\Xi(z)\phi_0^\Xi(z)dz$$

$$= \int_{\Gamma_\infty\backslash\mathbb{H}} \chi_T(z)\phi_0^\Xi(z)dz$$

$$= \frac{1}{\sqrt{[\Gamma : \Xi]}}\int_{1/T}^\infty \left(\int_N \phi_0^\Gamma(z)dx\right) \frac{dy}{y^2}, \qquad (4.13)$$

where N is a finite interval corresponding to the width of the parabolic at infinity in $\Gamma\backslash\mathbb{H}$. The key step here is that we have assumed $\Gamma_\infty \subset \Xi \subset \Gamma$, so Ξ contains the same maximal unipotent subgroup as Γ. Since Γ_∞ is assumed to be nontrivial, the Hausdorff dimension of the limit set of Γ satisfies

$$\delta > 1/2.$$

Therefore ϕ_0 is an eigenfunction of the Laplacian Δ with eigenvalue $\lambda_0 = \delta(1-\delta)$, the constant term in its Fourier expansion inherits the same differential equation. Therefore the inner integral in (4.13) is $\alpha y^\delta + \beta y^{1-\delta}$, where α and β are some periods (constants depending on ϕ_0). Inserting this into (4.13) and computing the elementary integral,

we have

$$
\begin{aligned}
\widehat{F^\Xi_T}(\lambda_0) &= \frac{1}{\sqrt{[\Gamma : \Xi]}} \left(\frac{\beta}{\delta} T^\delta + \frac{\alpha}{1-\delta} T^{1-\delta} \right) \\
&= \frac{c'}{\sqrt{[\Gamma : \Xi]}} T^\delta + O(T^{1/2}),
\end{aligned} \tag{4.14}
$$

where $c' = \beta/\delta > 0$ since ϕ_0 is nonnegative. Note that c' only depends on Γ and not on Ξ or ε (which has been suppressed until now).

Returning to (4.11), use (4.12) to define $c_{\Gamma,\varepsilon}$ by

$$
\frac{c_{\Gamma,\varepsilon}}{[\Gamma : \Xi]} = \frac{c'}{\sqrt{[\Gamma : \Xi]}} \left\langle \Psi^\Xi_{\varepsilon,\gamma z_0}, \phi_0^\Xi \right\rangle_\Xi = \frac{c'}{[\Gamma : \Xi]} \left\langle \Psi^\Xi_{\varepsilon,\gamma z_0}, \phi_0^\Gamma \right\rangle_\Xi, \tag{4.15}
$$

as ϕ_0 is Ξ-invariant. Using (4.15), we see that the main term in (4.9) coincides with the corresponding term in (4.4).

For more precise information on $c_{\Gamma,\varepsilon}$, in particular its dependence on ε, use the Mean Value Theorem and (4.12) to get

$$
\begin{aligned}
\left\langle \Psi^\Xi_{\varepsilon,\gamma z_0}, \phi_0^\Gamma \right\rangle_\Xi &= \int_{\Xi\backslash\mathbb{H}} \Psi^\Xi_{\varepsilon,\gamma z_0}(z) \phi_0^\Gamma(\gamma z_0) dz \tag{4.16} \\
&\quad + \int_{\Xi\backslash\mathbb{H}} \Psi^\Xi_{\varepsilon,\gamma z_0}(z) \left(\phi_0^\Gamma(z) - \phi_0^\Gamma(\gamma z_0) \right) dz \\
&= \phi_0^\Gamma(\gamma z_0) + O\left(\sup_{w \in B_\varepsilon(\gamma z_0)} \phi_0'(w) \cdot \varepsilon \right) \\
&= \phi_0^\Gamma(z_0) + O(\varepsilon), \tag{4.17}
\end{aligned}
$$

since ϕ_0^Γ is Γ-automorphic and $\int \Psi^\Xi = 1$. Thus (4.17), together with (4.15), verifies (4.5), where explicitly,

$$
c_\Gamma = c' \phi_0^\Gamma(\mathrm{i}).
$$

Returning to the rest of the spectrum in (4.9), we apply the Main Identity (3.8) (valid for arbitrary Γ, in particular for $\Gamma = \Xi$) to the error term:

$$
\begin{aligned}
Err &= \int_{\mathrm{Spec}(\Xi\backslash\mathbb{H})-\lambda_0} \widehat{F^\Xi_T}(\lambda) \widehat{\Psi}^\Xi_{\varepsilon,\gamma z_0}(\lambda) d\nu_\Xi(\lambda) \tag{4.18} \\
&= \int_{\mathrm{Spec}(\Xi\backslash\mathbb{H})-\lambda_0} \widehat{K_T(\Delta)} F^\Xi_1(\lambda) \widehat{\Psi}^\Xi_{\varepsilon,\gamma z_0}(\lambda) d\nu_\Xi(\lambda) \\
&\quad + \int_{\mathrm{Spec}(\Xi\backslash\mathbb{H})-\lambda_0} \widehat{L_T(\Delta)} F^\Xi_b(\lambda) \widehat{\Psi}^\Xi_{\varepsilon,\gamma z_0}(\lambda) d\nu_\Xi(\lambda),
\end{aligned}
$$

where $b < 3$. Assume $\lambda_1 < 1/4$. By (2.5), $\widehat{K_T(\Delta)F_1^\Xi}(\lambda) = K_T(\lambda)\widehat{F_1^\Xi}(\lambda)$, so together with Cauchy-Schwarz and the bound (3.6) we have

$$\int_{\mathrm{Spec}(\Xi\backslash\mathbb{H})-\lambda_0} \widehat{K_T(\Delta)F_1^\Xi}(\lambda)\widehat{\Psi}_{\varepsilon,\gamma z_0}^\Xi(\lambda)d\nu_\Xi(\lambda) \ll T^{s_1}\|F_1^\Xi\|_\Xi\|\Psi_\varepsilon^\Xi\|_\Xi,$$

(4.19)

where $\|\cdot\|_\Xi$ is the norm on $L^2(\Xi\backslash\mathbb{H})$. Clearly the inclusion $\Xi \subset \Gamma$ and positivity of χ_T implies the pointwise bound

$$F_T^\Xi(z) = \sum_{\xi\in\Gamma_\infty\backslash\Xi} \chi_T(\xi z) \leq \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \chi_T(\gamma z) = F_T^\Gamma(z).$$

Applying the pointwise bound directly gives

$$\|F_T^\Xi\|_\Xi \leq \|F_T^\Gamma\|_\Xi = \sqrt{[\Gamma:\Xi]}\|F_T^\Gamma\|_\Gamma,$$

but this is not good enough for us – we will lose information in the sieve! Instead, we can exploit the positivity of F_T^Ξ to unfold the L^2 norm with respect to one copy of F_T^Ξ, apply the pointwise bound to the other copy, and refold again:

$$
\begin{aligned}
\|F_T^\Xi\|_\Xi^2 &= \int_{\Xi\backslash\mathbb{H}} \left|F_T^\Xi(z)\right|^2 dz = \int_{\Xi\backslash\mathbb{H}} \left(\sum_{\xi\in\Gamma_\infty\backslash\Xi} \chi_T(\xi z)\right) F_T^\Xi(z)dz \\
&= \int_{\Gamma_\infty\backslash\mathbb{H}} \chi_T(z)F_T^\Xi(z)dz \\
&\leq \int_{\Gamma_\infty\backslash\mathbb{H}} \chi_T(z)F_T^\Gamma(z)dz = \int_{\Gamma\backslash\mathbb{H}} \left(\sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \chi_T(\gamma z)\right) F_T^\Gamma(z)dz \\
&= \|F_T^\Gamma\|_\Gamma^2,
\end{aligned}
$$

(4.20)

this time losing no information! Note that we have again used crucially the fact that $\Gamma_\infty \subset \Xi \subset \Gamma$.

As Ψ_ε^Ξ is an ε-approximation to the identity on a two-dimensional space, we can choose it so that

$$\|\Psi_\varepsilon^\Xi\|_\Xi \ll \frac{1}{\varepsilon},$$

(4.21)

where the implied constant is independent of Ξ. Combining (4.20) with (4.21), inserting into (4.19), and carrying out the same computation with L_T replacing K_T, we have

$$(4.18) \ll \frac{1}{\varepsilon}T^{s_1}\|F_3^\Gamma\|_\Gamma \ll \frac{1}{\varepsilon}T^{s_1},$$

since $b < 3$. The case $\lambda_1 \geq 1/4$ is similar.

This completes the proof of the Lemma. $\qquad\square$

The method of proof allows a much more general statement. Instead of pulling out only the contribution from the base eigenvalue λ_0 in (4.9), we can take more terms. Let $\theta \in (1/2, \delta)$ be a fixed constant, and denote the eigenvalues in $\mathrm{Spec}(\Xi\backslash\mathbb{H})$ below $\theta(1-\theta)$ by

$$\delta(1-\delta) = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M < \theta(1-\theta). \qquad (4.22)$$

Corresponding to each point eigenvalue λ_j is a normalized L^2 eigenfunction, ϕ_j. It may be the case that ϕ_j (which a priori is only Ξ-automorphic) is also automorphic with respect to some group Γ_j which satisfies $\Xi \subset \Gamma_j \subset \Gamma$. Then ϕ_j is an "oldform" on Ξ and $\phi_j \in L^2(\Gamma_j\backslash\mathbb{H})$. This means $\lambda_j = s_j(1 - s_j) \in \mathrm{Spec}(\Gamma_j\backslash\mathbb{H})$ and the same analysis as above gives

$$\widehat{F}_T^\Xi(\lambda_j)\widehat{\Psi}^\Xi(\lambda_j) = \frac{c_j T^{s_j}}{[\Gamma_j : \Xi]} + O(T^{1/2}),$$

where c_j depends on Γ_j but not on Ξ. After extracting these lower order terms, the remaining error is simply $O(\frac{1}{\varepsilon}T^\theta)$. Suppressing the precise dependence on ε (which will be fixed for the remainder of this section), we have proved

Lemma 4.2. *Let $\Gamma_\infty \subset \Xi \subset \Gamma$ with $[\Gamma : \Xi] < \infty$ and w_T be defined by (4.2). Fix $\theta \in (1/2, \delta)$ and let the eigenvalues in $\mathrm{Spec}(\Xi\backslash\mathbb{H})$ below $\theta(1-\theta)$ be denoted as in (4.22). For each $j = 1, \ldots, M$ let Γ_j denote a group satisfying $\Xi \subset \Gamma_j \subset \Gamma$ such that $\lambda_j \in \mathrm{Spec}(\Gamma_j\backslash\mathbb{H})$.*

Then for any fixed $\gamma \in \Gamma$, there exists a constant $c_\Gamma > 0$ depending only on Γ and ε, and constants c_j depending on Γ_j and ε, such that

$$\sum_{\xi \in \Gamma_\infty\backslash\Xi} w_T(\xi\gamma^{-1}) = \frac{c_\Gamma T^\delta}{[\Gamma : \Xi]} + \sum_{j=1}^M \frac{c_j T^{s_j}}{[\Gamma_j : \Xi]} + O_\varepsilon\left(T^\theta\right). \qquad (4.23)$$

4.2. **Sums over a_n.** Recall the notation $f(\gamma) = c^2 + d^2$ for a matrix $\gamma \in \Gamma$ having bottom row (c, d). For $n \geq 1$ let

$$a_n(T) := \sum_{\substack{\gamma \in \Gamma_\infty\backslash\Gamma \\ f(\gamma)=n}} w_T(\gamma) \qquad (4.24)$$

be a smoothed count for the number of elements in our orbit having height bounded by T and f-value exactly equal to n.

Recall that $\Gamma_\infty \subset \Gamma$ is the group which stabilizes $f(\gamma) = c^2 + d^2$ in the sense that for $\gamma' \in \Gamma_\infty$, we have $f(\gamma'\gamma) = f(\gamma)$. Recall from (2.2) the principal "congruence" subgroup of level q

$$\Gamma(q) := \{\gamma \in \Gamma : \gamma \equiv I(q)\}.$$

This group stabilizes all $\gamma \bmod q$, i.e. if $\gamma' \in \Gamma(q)$ then $\gamma'\gamma \equiv \gamma(q)$. Similarly, let

$$\Gamma_1(q) := \{\gamma \in \Gamma : \gamma \equiv \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} (\bmod q)\} \qquad (4.25)$$

be the subgroup of Γ which stabilizes $f(\bmod q)$, i.e. if $g \in \Gamma_1(q)$ then $f(g\gamma) \equiv f(\gamma)(q)$. The inclusions $\Gamma_\infty \subset \Gamma_1(q)$ and $\Gamma(q) \subset \Gamma_1(q) \subset \Gamma$ are immediate. In particular, $\mathrm{Spec}(\Gamma_1(q)\backslash\mathbb{H}) \subset \mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})$ and $\Gamma_1(q)$ inherits the spectral gap properties of $\Gamma(q)$.

Fix $q \geq 1$ square-free and consider

$$\sum_{n\equiv 0(q)} a_n(T). \qquad (4.26)$$

Insert (4.24) into (4.26) and decompose $\gamma \in \Gamma_\infty\backslash\Gamma$ into $\gamma = \xi\gamma_1$ with $\xi \in \Gamma_\infty\backslash\Gamma_1(q)$ and $\gamma_1 \in \Gamma_1(q)\backslash\Gamma$:

$$\sum_{n\equiv 0(q)} a_n(T) = \sum_{\substack{\gamma\in\Gamma_\infty\backslash\Gamma \\ f(\gamma)\equiv 0(q)}} w_T(\gamma) = \sum_{\gamma_1\in\Gamma_1(q)\backslash\Gamma} \sum_{\substack{\xi\in\Gamma_\infty\backslash\Gamma_1(q) \\ f(\xi\gamma_1)\equiv 0(q)}} w_T(\xi\gamma_1)$$

$$= \sum_{\substack{\gamma_1\in\Gamma_1(q)\backslash\Gamma \\ f(\gamma_1)\equiv 0(q)}} \left(\sum_{\xi\in\Gamma_\infty\backslash\Gamma_1(q)} w_T(\xi\gamma_1) \right), \qquad (4.27)$$

since $f \bmod q$ is invariant under $\Gamma_1(q)$. Apply Lemma 4.2 with $\Xi = \Gamma_1(q)$ to the inner sum in (4.27) to prove

Proposition 4.3. *Let $T > 1$, $q \geq 1$ be square-free and $a_n(T)$ be defined by (4.24). Let θ be the spectral gap of Γ and let the eigenvalues in $\mathrm{Spec}(\Gamma_1(q)\backslash\mathbb{H})$ below $\theta(1-\theta)$ be denoted by*

$$\delta(1-\delta) = \lambda_0 < \lambda_1(q) \leq \lambda_2(q) \leq \cdots \leq \lambda_{M(q)}(q) < \theta(1-\theta).$$

For each $j = 1, \ldots, M(q)$ let q_j denote a divisor $q_j | q$ such that

$$\lambda_j(q) = s_j(q)(1 - s_j(q)) \in \mathrm{Spec}(\Gamma_1(q_j)\backslash\mathbb{H}).$$

Then there exists a constant $c_\Gamma > 0$ depending on Γ and ε, and constants c_j depending on q_j, such that

$$\sum_{n\equiv 0(q)} a_n(T) = |\mathcal{O}_q| \left(\frac{c_\Gamma T^\delta}{[\Gamma : \Gamma_1(q)]} + \sum_{j=1}^{M(q)} \frac{c_j T^{s_j(q)}}{[\Gamma_1(q_j) : \Gamma_1(q)]} + O\left(\frac{1}{\varepsilon}T^\theta\right) \right),$$

$$(4.28)$$

where

$$|\mathcal{O}_q| = |\{(c,d)(\bmod q) \in \mathcal{O} \mid c^2 + d^2 \equiv 0(q)\}| = \sum_{\substack{\gamma_1\in\Gamma_1(q)\backslash\Gamma \\ f(\gamma_1)\equiv 0(q)}} 1. \qquad (4.29)$$

5. Proof of the Main Theorem

In this section we prove Theorem 1.3.

5.1. **Part (1).** We begin by proving (1.5). Let

$$H_\varepsilon(T) := \sum_n a_n(T),$$

with $a_n(T)$ defined in (4.24) and implicitly dependent on ε via $w_{T,\varepsilon}$ in (4.2). Recall the definition of $\mathcal{O}(T)$ from (1.4). By (4.3) we have

$$|\mathcal{O}(T/(1+\varepsilon))| \le H_\varepsilon(T) \le |\mathcal{O}(T/(1-\varepsilon))|,$$

or equivalently,

$$H_\varepsilon(T(1-\varepsilon)) \le |\mathcal{O}(T)| \le H_\varepsilon(T(1+\varepsilon)). \tag{5.1}$$

Apply Lemma 4.1 with $\Xi = \Gamma$ and $\gamma = I$:

$$H_\varepsilon(T) = (c_\Gamma + O(\varepsilon))T^\delta + O(\frac{1}{\varepsilon}T^{s_1}), \tag{5.2}$$

where $\lambda_1 = s_1(1-s_1)$ is the first eigenvalue above the base in $\mathrm{Spec}(\Gamma\backslash\mathbb{H})$, satisfying $\delta > s_1$. Combining (5.1) with (5.2) we elementarily arrive at (1.5) by an appropriate choice of ε.

5.2. **Part (2).** Our only remaining task is to verify all of the conditions necessary to apply a combinatorial sieve to $\mathcal{A} = \{a_n(T)\}$ (see §2.3). By (4.3), the sequence $a_n(T)$ defined by (4.24) satisfies (2.9) with $T(1+\varepsilon)$ replacing T. Throughout the rest of the section, ε will be a fixed small constant. Anyway since the bound on R in (2.11) depends on $\log T$, this difference is irrelevant.

Let $\theta \in [1/2, \delta)$ be a spectral gap for Γ. There are at most finitely many primes for which Strong Approximation fails, and also finitely many primes at which the corresponding spectrum fails to have a θ-gap. Let $\mathfrak{B}$ be the product of the primes in these finite "ramified" sets.

For ease of exposition, assume first that $q \ge 1$ is square-free and relatively prime to $\mathfrak{B}$. Then the projection

$$\Gamma \longrightarrow \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z}) \text{ is onto}, \tag{5.3}$$

and

$$\mathrm{Spec}(\Gamma(q)\backslash\mathbb{H}) \cap (0, \theta(1-\theta)) = \mathrm{Spec}(\Gamma\backslash\mathbb{H}) \cap (0, \theta(1-\theta)). \tag{5.4}$$

Apply Proposition 4.3 to (2.8), and infer from (5.4) that we can take $q_j = 1$ for all j:

$$|\mathcal{A}_q| = |\mathcal{O}_q| \left(\frac{1}{[\Gamma : \Gamma_1(q)]} \left(c_\Gamma T^\delta + \sum_j c_j T^{s_j} \right) + O_\varepsilon(T^\theta) \right),$$

where all c_j's only depend on Γ (and ε, which is fixed) but not on q. Then $|\mathcal{A}_q| = \omega(q)\mathcal{X} + r(q)$, with

$$\omega(q) = \frac{|\mathcal{O}_q|}{[\Gamma : \Gamma_1(q)]},$$
$$\mathcal{X} = c_\Gamma T^\delta + \sum_j c_j T^{s_j},$$

and

$$r(q) = O(|\mathcal{O}_q|T^\theta). \tag{5.5}$$

Trivially $\omega(1) = 1$. By (4.29) and (5.3) it is clear that ω is multiplicative, and for $q = p$ a prime we can compute $\omega(p)$ exactly.

As p is unfamified, $|\mathcal{O}_p|$ counts the number of $(c, d) \bmod p$ with $(c, d) \neq (0, 0)$ and $c^2 + d^2 \equiv 0(p)$. The last equation is equivalent to $c^2 \equiv -d^2(p)$, which has no solutions if -1 is not a square mod p, that is, when $p \equiv 3(4)$. In the opposite case, this cardinality is easily computed by hand. Similarly, $[\Gamma : \Gamma_1(p)]$ counts the number of $(c, d) \bmod p$ with $(c, d) \neq (0, 0)$. Thus we have

$$|\mathcal{O}_p| = \begin{cases} 1 & \text{if } p = 2 \\ 2(p-1) & \text{if } p \equiv 1(4) \\ 0 & \text{if } p \equiv 3(4), \end{cases} \tag{5.6}$$
$$[\Gamma : \Gamma_1(p)] = p^2 - 1,$$

and so

$$\omega(p) = \frac{|\mathcal{O}_p|}{[\Gamma : \Gamma_1(p)]} = \begin{cases} 1/3 & \text{if } p = 2 \\ 2/(p+1) & \text{if } p \equiv 1(4) \\ 0 & \text{if } p \equiv 3(4). \end{cases}$$

Then the sieve condition (S2) is obvious, and (S3) follows from

$$\prod_{\substack{p < z \\ p \equiv 1(4)}} \left(1 - \frac{2}{p} \right)^{-1} \sim \kappa \log z,$$

a classical exercise (see e.g. [Lan53]). Inserting (5.6) into (5.5) gives

$$r(q) = O(qT^\theta), \tag{5.7}$$

and since $\mathcal{X} \asymp T^\delta$, (S5) requires

$$\sum_{q \leq Q} |r(q)| = O(Q^2 T^\theta) \ll T^{\delta/(1+\epsilon)},$$

for any $\epsilon > 0$. This is satisfied for

$$Q = T^{\frac{(\delta-\theta)}{2(1+\epsilon)}}. \tag{5.8}$$

Then inputting (5.8) into (2.11), together with (2.10), gives (1.6) for

$$R > 2\log T/\log Q = 4/(\delta - \theta),$$

since $\epsilon > 0$ was arbitrary.

This completes the analysis of the affine linear sieve in the case q is "unramified".

5.3. **Ramified places.** Let $q = q'q'' \geq 1$ be square-free, with $q'|\mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$. Notice that $\mathfrak{B}$ being the product of a finite number of primes means there are only finitely many possible values of q'.

Then by Theorem 2.1 the projection of Γ in $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ is

$$G_{q'} \times \mathrm{SL}_2(\mathbb{Z}/q''\mathbb{Z}), \tag{5.9}$$

where $G_{q'}$ is the projection of Γ in $\mathrm{SL}_2(\mathbb{Z}/q'\mathbb{Z})$.

By Theorem 2.3, if $\lambda < \theta(1 - \theta)$ is in $\mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})$ but not in $\mathrm{Spec}(\Gamma\backslash\mathbb{H})$, then $\lambda \in \mathrm{Spec}(\Gamma(q')\backslash\mathbb{H})$.

Then applying Proposition 4.3 and using the fact that

$$[\Gamma : \Xi] = [\Gamma : \Gamma_j][\Gamma_j : \Xi]$$

for $\Xi \subset \Gamma_j \subset \Gamma$, we have

$$|\mathcal{A}_q| = |\mathcal{O}_q| \left(\frac{1}{[\Gamma : \Gamma_1(q)]} (\mathcal{X} + \mathcal{X}_q) + O_\varepsilon(T^\theta) \right),$$

where

$$\mathcal{X} = c_\Gamma T^\delta + \sum_{\substack{\lambda_j \in \mathrm{Spec}(\Gamma\backslash\mathbb{H}) - \{\lambda_0\} \\ \lambda_j = s_j(1-s_j) < \theta(1-\theta)}} c_j T^{s_j},$$

and

$$\mathcal{X}_{q'} = [\Gamma : \Gamma_1(q')] \sum_{\substack{\lambda_j^* \in \mathrm{Spec}(\Gamma_1(q')\backslash\mathbb{H})_{new} \\ \lambda_j^* = s_j^*(1-s_j^*) < \theta(1-\theta)}} c_j^* T^{s_j^*}.$$

Here c_j's depend on Γ and not on q, while c_j^*'s depend on q' but not on q''. Clearly $\mathcal{X}_{q'} \ll \mathcal{X}^{1-\eta}$ for some $\eta > 0$. By (5.9), $|\mathcal{O}_q| = |\mathcal{O}_{q'}||\mathcal{O}_{q''}|$ and $[\Gamma_1(q'') : \Gamma_1(q)] = [\Gamma : \Gamma_1(q')]$, which is independent of q''.

Therefore we have the expression

$$|\mathcal{A}_q| = \omega(q)(\mathcal{X} + \mathcal{X}_{q'}) + r(q),$$

where

$$\omega(q) = \frac{|\mathcal{O}_q|}{[\Gamma : \Gamma_1(q)]} = \frac{|\mathcal{O}_{q'}|}{[\Gamma : \Gamma_1(q')]} \frac{|\mathcal{O}_{q''}|}{[\Gamma : \Gamma_1(q'')]} = \omega(q')\omega(q''),$$

and $r(q)$ satisfies (5.7). The rest of the analysis follows as before, completing the proof of the Main Theorem.

APPENDIX A. PROOF OF THEOREM 2.5

In this appendix we derive Theorem 2.5 from the more standard sieve setting. See [Iwa96, IK04].

As before, let $\mathcal{A} = \{a_n\}$ be our sequence of nonnegative numbers with $a_n = 0$ for n exceeding a parameter T, and let $\mathfrak{B}$ be the product of a finite set of "bad" primes. Recall that $\mathcal{X}$ is an approximation to $|\mathcal{A}|$:

$$\mathcal{X} \asymp \sum_n a_n \asymp T^\delta$$

and that for $q = q'q''$ square-free with $q'|\mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$ we have

$$|\mathcal{A}_q| = \sum_{n \equiv 0(q)} a_n = \omega(q)(\mathcal{X} + \mathcal{X}_{q'}) + r(q).$$

It is assumed that for any divisor $q'|\mathfrak{B}$, the factor $\mathcal{X}_{q'}$ is a power less than the main term $\mathcal{X}$. Also ω is multiplicative away from $\mathfrak{B}$, i.e. $\omega(q) = \omega(q') \prod_{p|q''} \omega(p)$. Let $z \geq 2$ be a parameter (a small power of $\mathcal{X}$) and define

$$P(z) \ := \ \prod_{p<z} p.$$

Assume z is large enough (by taking T large enough) so that $\mathfrak{B}|P(z)$. Consider a sum of the form

$$S(z) \ := \ \sum_{(n,P(z))=1} a_n.$$

Notice that if $z = T^\alpha$, $n \leq T$ and $(n, P(z)) = 1$ then n has at most $1/\alpha$ prime factors. In this way, $S(z)$ counts the number of $1/\alpha$-almost primes.

By Möbius inversion, we have

$$
\begin{aligned}
S(z) &= \sum_n a_n \sum_{q|(n,P(z))} \mu(q) \\
&= \sum_{q|P(z)} \mu(q) \sum_{n \equiv 0(q)} a_n \\
&= \sum_{q|P(z)} \mu(q) \left(\omega(q)(\mathcal{X} + \mathcal{X}_{q'}) + r(q) \right) \\
&= \Sigma_1 + \Sigma_2 + \Sigma_3,
\end{aligned}
$$

with

$$
\Sigma_1 = \sum_{q|P(z)} \mu(q)\omega(q)\mathcal{X} = \mathcal{X} \prod_{\substack{p<z \\ p \nmid \mathfrak{B}}}(1 - \omega(p)) \times \sum_{q'|\mathfrak{B}} \mu(q')\omega(q'),
$$

$$
\Sigma_2 = \sum_{q|P(z)} \mu(q)\omega(q)\mathcal{X}_{q'} = \prod_{\substack{p<z \\ p \nmid \mathfrak{B}}}(1 - \omega(p)) \times \left(\sum_{q'|\mathfrak{B}} \mu(q')\omega(q')\mathcal{X}_{q'} \right),
$$

and

$$
\Sigma_3 = \sum_{q|P(z)} \mu(q)r(q).
$$

Let

$$
S^*(z) = \Sigma_1 + \Sigma_3.
$$

Theorem A.1 (See Iwaniec [Iwa96], Theorem 3.7 on page 63). *Let $Q > e^{2K}$ where K is the constant appearing in (3) and define*

$$
\begin{aligned}
V(z) &= \prod_{\substack{p \leq z \\ p \nmid \mathfrak{B}}}(1 - \omega(p)), \\
R(Q) &= \sum_{q \leq Q} |r(q)| \\
f(s) &= 2e^\gamma \log(s-1)/s, \text{ for } s \in [2,4], \\
F(s) &= 2e^\gamma/s, \text{ for } s \in [1,3], \text{ and} \\
D &= cK^{11}(\log\log\log Q)^3(\log\log Q)^{-1}.
\end{aligned}
$$

Here $\gamma = .577\ldots$ is the Euler constant and $c = 3.591\ldots$ solves $(c/e)^c = e$. Then

$$
(f(s) - D)\mathcal{X}V(z) - R(Q) \leq S^*(z) \leq (F(s) + D)\mathcal{X}V(z) + R(Q),
$$

where $s = \log Q/\log z$.

The choice $Q = T^{\frac{\delta - \theta}{2(1+\epsilon)}}$ gives

$$D \ll (\log\log\log T)^3 (\log\log T)^{-1} \to 0,$$

so we can take

$$s = 2(1 + \epsilon), \tag{A.1}$$

giving $f(s) > D$ for T sufficiently large.

Then with $z = Q^{1/s} = T^{\frac{\delta - \theta}{4(1+\epsilon)^2}}$ we have

$$S^*(z) \asymp \frac{\mathcal{X}}{\log \mathcal{X}} \asymp T^\delta / \log T \asymp S(z),$$

since $\Sigma_2 \ll \mathcal{X}^{1-\eta}$. As $\epsilon > 0$ is arbitrary, this captures R-almost primes with any $R > 4/(\delta - \theta)$.

Remark A.2. The set Z_R is genuinely the set of integers having at most R prime factors, not just R prime factors outside of $\mathfrak{B}$.

APPENDIX B. PROOF OF LEMMA 3.3

We require the following simple lemma from the theory of inhomogeneous ODEs, in particular the method of variation of parameters.

Lemma B.1. *Let* $\lambda \geq 0$ *and suppose* f *and* g *are functions satisfying*

$$-y^2 \frac{\partial^2}{\partial y^2} f(y) - \lambda f(y) = g(y).$$

Assume $\lambda = s(1 - s) \neq 1/4$. *Then there exist constants* α *and* β *such that*

$$f(y) = \alpha y^s + \beta y^{1-s} + u(y)y^s + v(y)y^{1-s},$$

where

$$u(y) = (1 - 2s)^{-1} \int_{1/T}^y w^{-1-s} g(w)\, dw \tag{B.1}$$

and

$$v(y) = (2s - 1)^{-1} \int_{1/T}^y w^{s-2} g(w)\, dw.$$

If $\lambda = 1/4$ *then*

$$f(y) = \alpha y^{1/2} + \beta y^{1/2} \log y + u(y)y^{1/2} + v(y)y^{1/2} \log y,$$

where

$$u(y) = \int_{1/T}^y w^{-3/2} \log(w) g(w)\, dw, \text{ and}$$

$$v(y) = -\int_{1/T}^y w^{-3/2} g(w)\, dw.$$

Proof. Elementary calculus. For details see [Kon07].   □

Fix any $\phi \in L^2(\Gamma \backslash \mathbb{H})$ and any $\lambda \geq 0$. For simplicity assume $\lambda \neq 1/4$ (the calculation in the opposite case is similar). Consider the left hand side of (3.9) and unfold:

$$
\begin{aligned}
\langle F_T, \phi \rangle &= \int_{\Gamma \backslash \mathbb{H}} F_T(z) \phi(z) dz \\
&= \int_{1/T}^{\infty} \left(\int_N \bar{\phi}(z) dx \right) \frac{dy}{y^2},
\end{aligned}
$$

where again N is an interval corresponding to the width of the cusp at infinity.

Let

$$
f(y) = \int_N \bar{\phi}(z) dx,
$$

so that

$$
\langle F_T, \phi \rangle = \int_{1/T}^{\infty} f(y) \frac{dy}{y^2}. \tag{B.2}
$$

Let g be defined by:

$$
g(y) := -y^2 \frac{\partial^2}{\partial y^2} f(y) - \lambda f(y) = \int_N (\Delta - \lambda) \phi(z) dx. \tag{B.3}
$$

By Lemma B.1,

$$
f(y) = \alpha y^s + \beta y^{1-s} + y^s u(y) + y^{1-s} v(y).
$$

The first two terms are the homogenous solution and the last two are the perturbation. Of course inserting the homogenous component of f into (B.2) we have the main term in (3.9)

$$
\int_{1/T}^{\infty} \left(\alpha y^s + \beta y^{1-s} \right) \frac{dy}{y^2} = AT^s + BT^{1-s}.
$$

Thus it remains to show that $I, II \ll_{\lambda, T} \|(\Delta - \lambda) \phi\|$, where

$$
I = \int_{1/T}^{\infty} y^s u(y) \frac{dy}{y^2}, \text{ and } II = \int_{1/T}^{\infty} y^{1-s} v(y) \frac{dy}{y^2}
$$

are the contributions from the perturbation. Integrate I by parts and recall from (B.1) that $u(1/T) = 0$:

$$
I = u(y) \frac{y^{s-1}}{s-1} \Big|_{y \to \infty} - \int_{1/T}^{\infty} \frac{y^{s-1}}{s-1} \left(\frac{y^{-1-s} g(y)}{1 - 2s} \right) dy. \tag{B.4}
$$

Modulo constants, insert (B.3) into the last integral of (B.4) and apply Cauchy-Shwarz:

$$\int_{1/T}^{\infty} \int_N \left(\frac{(\Delta - \lambda)\phi(z)}{y} \right) \left(\frac{1}{y} \right) dx\, dy$$

$$\ll \left(\int_{1/T}^{\infty} \int_N \left| \frac{(\Delta - \lambda)\phi(z)}{y} \right|^2 dx\, dy \right)^{1/2} \left(\int_{1/T}^{\infty} \int_N \left| \frac{1}{y} \right|^2 dx\, dy \right)^{1/2}$$

$$\ll_{\lambda, T, \Gamma} \quad \|(\Delta - \lambda)\phi\| \, (|N|\, T)^{1/2} \tag{B.5}$$

$$\ll_{\lambda, T, \Gamma} \quad \|(\Delta - \lambda)\phi\|,$$

where $|N|$ is the length of N. Here we used the fact that the box

$$N \times [1/T, \infty] \subset \mathbb{H}$$

is contained in a union of finitely many (depending on T) fundamental domains for $\Gamma \backslash \mathbb{H}$.

Remark B.2. The appearance of $|N|$ in (B.5) (via the use of Cauchy-Shwarz) is the **most severe** obstruction to removing the assumption that Γ stabilize infinity. If there is no cusp at infinity, then $|N| = \infty$ and our analysis fails. See Remark 1.7.

For the first part of (B.4), we need the following bound:

$$u(y) = \frac{1}{1 - 2s} \int_{1/T}^{y} w^{-1-s} g(w) dw$$

$$= \frac{1}{1 - 2s} \int_{1/T}^{y} \int_N \frac{(\Delta - \lambda)\phi(x + iw)}{w} \cdot w^{-s} dx\, dw$$

$$\ll_{\lambda} \left(\int_{1/T}^{y} \int_N \left| \frac{(\Delta - \lambda)\phi(x + iw)}{w} \right|^2 dx\, dw \right)^{1/2}$$

$$\times \left(\int_{1/T}^{y} \int_N \left| w^{-s} \right|^2 dx\, dw \right)^{1/2}$$

$$\ll_{\lambda, T} \|(\Delta - \lambda)\phi\| \begin{cases} y^{1-2s} + T^{2s-1} & \text{if } s > 1/2 \\ \log y + \log T & \text{if } \mathfrak{Re}(s) = 1/2. \end{cases}$$

So

$$\lim_{y \to \infty} u(y) y^{s-1} \ll_{\lambda, T} \|(\Delta - \lambda)\phi\|.$$

The integral II is handled identically and we are done.

Note added in proof: In joint work with Hee Oh [KO08a], we circumvent the Main Identity to prove the Main Theorem without the assumption that Γ_∞ is nontrivial. Instead of the Main Identity, we prove the equidistribution of long horocycle flows on the unit tangent bundle of an infinite-volume Riemann surface of constant negative curvature, and then use this equidistribution to count. Furthermore, we replace the β sieve by the weighted sieve of Diamond-Halberstam-Richert [DHR88], which gives better numbers under nearly identical hypotheses, and execute the sieve for various other choices of f in (1.1). We also use these methods in [KO08b] to count the number of circles in an Apollonian packing of bounded curvature, and discuss various Diophantine properties of integral Apollonian packings.

References

[Bea68]   A. F. Beardon. The exponent of convergence of Poincaré series. *Proc. London Math. Soc. (3)*, 18:461–483, 1968.

[Bea83]   Alan F. Beardon. *The Geometry of Discrete Groups*, volume 91 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1983.

[BG07]   Jean Bourgain and Alex Gamburd. Uniform expansion bounds for Cayley graphs of $SL_2(\mathbf{F}_p)$, 2007. Preprint.

[BGS06]   Jean Bourgain, Alex Gamburd, and Peter Sarnak. Sieving and expanders. *C. R. Math. Acad. Sci. Paris*, 343(3):155–159, 2006.

[BGS08]   Jean Bourgain, Alex Gamburd, and Peter Sarnak. Affine linear sieve, expanders and sum-product, 2008. Preprint, available at http://www.math.princeton.edu/sarnak/.

[Bru19]   V. Brun. Le crible d'Eratosthéne et le theoréme de Goldbach. *C. R. Acad. Sci. Paris*, 168:544–546, 1919.

[Che73]   Jing Run Chen. On the representation of a larger even integer as the sum of a prime and the product of at most two primes. *Sci. Sinica*, 16:157–176, 1973.

[DHR88]   H. Diamond, H. Halberstam, and H.-E. Richert. Combinatorial sieves of dimension exceeding one. *J. Number Theory*, 28(3):306–346, 1988.

[DSV03]   G. Davidoff, P. Sarnak, and A. Valette. *Elementary Number Theory, Group Theory and Ramanujan Graphs*, volume 55 of *London Math. Soc., Student Text*. Cambridge University Press, 2003.

[EMV07]   M. Einsiedler, G. Margulis, and A. Venkatesh. Effective equidistribution for closed orbits of semisimple groups on homogeneous spaces, 2007. arXiv:0708.4040.

[FI98]   John Friedlander and Henryk Iwaniec. The polynomial X^2+Y^4 captures its primes. *Ann. of Math. (2)*, 148(3):945–1040, 1998.

[Gam02]   Alex Gamburd. On the spectral gap for infinite index "congruence" subgroups of $SL_2(\mathbf{Z})$. *Israel J. Math.*, 127:157–200, 2002.

[Hal63]   P.R. Halmos. What does the spectral theorem say? *The American Mathematical Monthly*, 70(3):241–247, 1963.

[HB01]   D. R. Heath-Brown. Primes represented by $x^3 + 2y^3$. *Acta Math.*, 186(1):1–84, 2001.

[IK04]   Henryk Iwaniec and Emmanuel Kowalski. *Analytic Number Theory*, volume 53 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2004.

[Iwa78]   Henryk Iwaniec. Almost-primes represented by quadratic polynomials. *Invent. Math.*, 47:171–188, 1978.

[Iwa96]   Henryk Iwaniec. *Sieve Methods*. Graduate Course, Rutgers, 1996. Unpublished Manuscript.

[KO08a]   A. Kontorovich and H. Oh. Almost prime pythagorean triples in thin orbits, 2008. Preprint.

[KO08b]   A. Kontorovich and H. Oh. Apollonian circle packings and horospherical flows on hyperbolic 3-manifolds, 2008. Preprint *http://arxiv.org/abs/0811.2236*.

[Kon07]   A. V. Kontorovich. *The Hyperbolic Lattice Point Count in Infinite Volume with Applications to Sieves*. Columbia University Thesis, 2007.

[KS03]   H. Kim and P. Sarnak. Refined estimates towards the Ramanujan and Selberg conjectures. *J. Mar. Math. Soc.*, 16:175–181, 2003.

[KS08]   D. Kelmer and P. Sarnak. Spectral gap for products of $PSL(2, \mathbb{R})$, 2008. Preprint, available at http://www.math.princeton.edu/sarnak/.

[Lan53]   Edmund Landau. *Handbuch der Lehre von der Verteilung der Primzahlen. 2 Bände*. Chelsea Publishing Co., New York, 1953. 2d ed, With an appendix by Paul T. Bateman.

[LP82]   P.D. Lax and R.S. Phillips. The asymptotic distribution of lattice points in Euclidean and non-Euclidean space. *Journal of Functional Analysis*, 46:280–350, 1982.

[MVW84]   C. Matthews, L. Vaserstein, and B. Weisfeiler. Congruence properties of Zariski-dense subgroups. *Proc. London Math. Soc*, 48:514–532, 1984.

[Pat76]   S.J. Patterson. The limit set of a Fuchsian group. *Acta Mathematica*, 136:241–273, 1976.

[Pu53]   I. I. Pjateckiĭ-Šapiro. On the distribution of prime numbers in sequences of the form $[f(n)]$. *Mat. Sb.*, 33:559–566, 1953.

[Sar07]   Peter Sarnak. Equidistribution and primes, 2007. PIMS Lecture, available at http://www.math.princeton.edu/sarnak/.

[Sel56]   A. Selberg. Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *J. Indian Math. Soc. (N.S.)*, 20:47–87, 1956.

[Sel65]   A. Selberg. On the estimation of Fourier coefficients of modular forms. *Proc. of Symposia in Pure Math.*, VII:1–15, 1965.

[Sul84]   D. Sullivan. Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups. *Acta Math.*, 153(3-4):259–277, 1984.

E-mail address: alexk@math.brown.edu

DEPARTMENT OF MATHEMATICS, COLUMBIA UNIVERSITY, 2990 BROADWAY, NEW YORK, NY 10027

Current address: Department of Mathematics, Brown University, 151 Thayer St, Providence, RI 02912